Language Generation
for Cross-Lingual Document Summarisation

Stephan Busemann

DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
busemann@dfki.de

Abstract: User-adaptive summaries of longer texts in the user's language are a major prerequisite for successful and efficient navigation in the results offered by WWW search engines and information retrieval systems. Current summarisation systems are either monolingual or use existing low-performance machine translation technology to target the user's desired language. What is more, their results cannot be tailored to the user's needs. The present contribution describes generation techniques required for cross-lingual summarisation, both with respect to the content of the summary and to user-oriented meta information about the original document. These techniques have been implemented in the MUSI system that summarises scientific medical texts originally written in Italian or English in German or French. MUSI uses deep linguistic processing on selected parts of the document. By using query-based selection and presenting additional information about the documents upon request, MUSI generates summaries tailored towards the user's needs.

Keywords: Language generation, cross-lingual summarisation, interlingua semantics representation, user-adaptive summarisation.

1 Introduction

User-adaptive summaries of longer texts in the user's language are a major prerequisite for successful and efficient navigation in the results offered by multi-lingual IR systems. Current summarisation systems, as found on WWW search engines, are either monolingual or use existing low-performance machine translation technology to target the user's desired language. What is more, their results cannot be tailored to the user's needs.

These shortcomings called for the design and development of a multi-lingual summarisation platform comprising systems based on different summarisers and language models. A major goal is to provide summaries in the user's desired language, irrespective of the language the original document was written in. The user should be able to tailor the summaries according to her needs.

The MUSI[1] summarisation system was designed to partially implement this goal. MUSI concentrates on only one approach to summarisation, on four different languages, and on a small set of scientific medical papers. Documents in Italian or English can be summarised in French and German. The user can influence the summaries by her query and by specifying types of information *about* the document that she is interested in, in order to support her assessment of document relevance. For instance, this may include the source language, the length of the document, or the location of the summary material within the document. Our major use case

[1] MUSI (Multilingual Summarisation for the Internet) is a project that was funded by the European Commission in 2000-2001. Partners are LexiQuest (Paris, France), Istituto Linguistica Computazionale (Pisa, Italy), and DFKI.

is indicative summarisation in document retrieval, i.e. deciding whether or not to download and read the full document. The primary goal of MUSI is to demonstrate the validity and the usability of the approach taken. It will be left to follow-up activities to realise the larger view of an application-oriented summarisation platform more completely.

The present contribution focuses on one part of the MUSI system, the human language generation techniques required for cross-lingual summarisation, both with respect to the content of the summary and to information *about* the original document. More precisely, this paper discusses the generation of German language summaries of scientific medical texts written in Italian or English.

MUSI has adopted an interlingua approach in order to make the contents of selected sentences amenable to the generation components. The interlingua was defined as a semantic representation language that could both be targeted by analysis components and define the input for generation. Being the fourth stage of internal representation in the system, it was simply termed IRep4(for Internal Representation 4). As will be discussed later, IRep4 expressions exhibit a high degree of granularity in order to account for the many linguistic distinctions encountered in the freely formulated texts of the genre investigated. In constrast, statements *about* the document – or meta-statements, as we will call them – can be based on much simpler grounds. Hence the generation task differs as well.

The system TG/2 [Busemann, 1996] was reused for the generation of German. TG/2 has been shown to be particularly well-suited to generating dialog contributions [Busemann *et al.*, 1994] and to report generation [Busemann and Horacek, 1998]. In either case, the input to the system was a non-linguistic representation of domain-semantic facts and actions. For MUSI, the generation of meta-statements was very much in line with these previous usages of TG/2, but the fine-grained IRep4 representations would expand the application space of TG/2 considerably.

The paper is organised as follows. Section 2 presents the MUSI system to the extent needed to discuss generation. This is followed by a sample summary of the type that can be generated by the system. Section 3 gives a closer look at IRep4 and the input for meta-statements. The generation techniques employed in MUSI are then presented in Section 4.

2 Cross-Lingual Summarisation in MUSI

In the field of text summarisation, statistics-based techniques are currently prevalent (cf. [Mani and Maybury, 1999]). They have the considerable advantage of not being tied to a particular coverage of grammars – in fact, the algorithms can, in principle, be applied to texts in any language – or to some domain, for which the content can be analysed reliably. Summarisation is seen mainly as a process of sentence extraction and concatenation. Among the drawbacks of purely statistical approaches, the difficulty of adapting the summaries to different user needs and the fact that cross-lingual summaries cannot be created are most noteworthy.

As an alternative to statistical approaches, rule-based techniques can overcome the drawbacks mentioned. Crossing the language barrier is possible with the representation of contents at a conceptual level. The possibility of including new text into the summary opens up opportunities for user-oriented summary formulation. Moreover, multi-document summarisation, which includes the need for information fusion, becomes possible [Radev and McKeown, 1998]. However, concept-based techniques suffer from other, well-known drawbacks, most prominently their dependence on domain and linguistic knowledge, and, as a consequence, their effective lack of scalability.

Both approaches can be combined in different ways. The way chosen in MUSI will now be described using the architectural overview shown in Figure 1. MUSI is based on a pipelined architecture of five modules linked up by intermediate representations. These are realised using XML and accumulate the results of previous components in the pipeline.

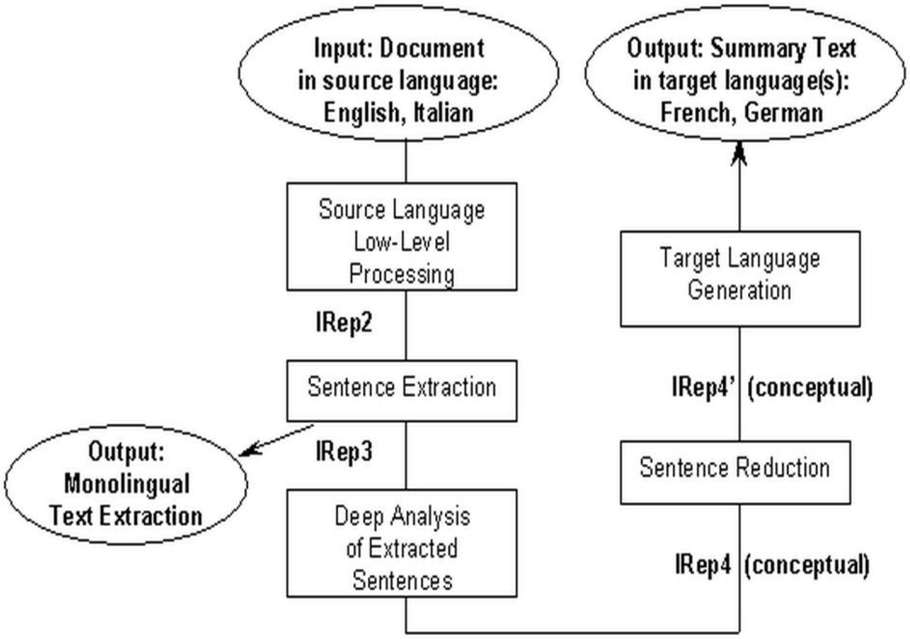

Input: Document in source language: English, Italian
Output: Summary Text in target language(s): French, German
Source Language Low-Level Processing
IRep2
Sentence Extraction
IRep3
Output: Monolingual Text Extraction
Deep Analysis of Extracted Sentences
Target Language Generation
IRep4' (conceptual)
Sentence Reduction
IRep4 (conceptual)

Figure 1: The MUSI System Architecture.

The input document is first analysed at a low level to identify sections, section headings and paragraphs. This information is then used by the extraction module, which filters the document using the query terms and the thesaurus-based expansions thereof, generic and domain-specific cue-phrases, and the position of the sentence within the document. A weight based on empirically determined constants is assigned to each sentence according to the following formula:

$$weight(S) = c_1 \cdot position(S) + c_2 \cdot cue\_phrases(S) + c_3 \cdot query\_relevance(S)$$

The value of $position()$ is higher if S occurs in the abstract or the conclusion rather than in the body of the document. The two other functions depend on whether S contains a cue phrase or an element of the set of search term expansions, respectively.

Different methods for selecting the sentences for the summary can be applied, among them the selection of any sentence with $weight(S) > t$, with t being some threshold, or simply the selection of the n top-ranked sentences of the document. This constitutes the statistics-based part of MUSI.

Note that by concatenating the results, an extract of the original document is gained that can serve as a monolingual summary.[2]

Each of the selected sentences is then subjected to deep linguistic analysis that yields an IRep4 expression. Since the analysis phase concentrates on syntax and linguistic semantics, ambiguities and unresolved relations can arise. MUSI does not attempt do resolve these, but resorts to heuristics – e.g. a pronoun usually refers to the closest compatible noun preceding it in the same sentence – or to splitting up the IRep4 structure into analysable fragments.

The IRep4 expressions may contain information that can be identified as irrelevant to the user query by virtue of the analysis results. Such pieces of information need not be reproduced during generation. The IRep4 is thus streamlined and shortened, and a filtering of fragmental results occurs.[3]

[2] No user-oriented tailoring has been provided for this, although technically it would have been possible.

[3] This module has only been realised in a very rudimentary form in the current implementation.

MUSI findet die folgende Information über das Originaldokument. Das Originaldokument ist in englisch geschrieben. Es ist in Bezug auf die Anfrage relativ relevant. Der Artikel ist 3454 Worte lang, er hat eine Tabelle, drei Abbildungen und eine Bibliographie. Alle Sätze der Zusammenfassung entstammen dem Abschnitt 'Diskussion'. Hier folgt die Zusammenfassung:

Ein kompetitiver Antagonismus zu Acetylcholin auf dem Niveau der muskarinischen Bindungsstellen dieser Substanzen verursacht die Wirkungen. Die Pflanze, die in dem vorliegenden Fall verzehrt worden ist, war Mandragora officinarium, die mit einer anderen Pflanze verwechselt wurde: die als Borretsch bekannte Borago Officinalis der Borretschgewächse, die wild auf Feldern wächst und in unserer Region gekocht als Gemüse gegessen wird. [...] die resultierende Akkumulation von Acetylcholin auf synaptischem Niveau neigt dazu, die anticholinergen Wirkungen der genannten Substanzen auszugleichen. [...] Intramuskuläre Verabreichung von Physostigmin und intravenöse anhaltende Infusionen werden wegen unberechenbarer Aufnahme nicht empfohlen.

[MUSI finds the following information about the source document. The source document is written in English. It is relatively relevant to the query. The article is 3454 words long, it has one table, three figures, and a bibliography. All sentences of the summary originate from the section 'Discussion'. Here is the summary:

The effects are caused by a competitive antagonism with acetylcholine on the level of the muscarinic sights of these substances. The plant ingested in the present case was mandragora officinarum of a species of potato which was mistaken for another plant: Borago officinalis of the boracic species known as borage, which grows wild in fields and is normally eaten cooked as a vegetable in our region. [...] the resulting accumulation of acetylcholine on the synaptic level tends to balance the anticolinergic effects of the substances cited. [...] Intramuscular administration of physostigmine is not recommended because of unpredictable absorption, as is prolonged intravenous infusion.]

Figure 2: A Sample German MUSI Summary from an English text. The meta-statements are translated into English. The English content part constitutes the monolingual extract).

The resulting representation(s) – called IRep4' in Figure 1 – constitute the input for the generation of the summary content. During the whole process, information *about* the document is accumulated as it becomes available; it serves as input for the generation of meta-statements.

The system comprises different components for analysis and generation. These components have been reused by the project partners and adapted towards the needs of IRep4. Their integration suggests that MUSI can, in fact, be expanded into a platform for configuring diverse summarisation systems. The need for different components was dictated by the different languages involved and by the pre-existing software each project partner would be able to introduce. A considerable effort was spent for the creation and adaptation of the linguistic resources. The approach chosen requires domain-specific lexicons that are based on a shared set of domain concepts used in IRep4. Besides, the domain-specific words had to be added to the morphological lexicons, so as to enable e.g. the generation of German word forms.

In general, query-based summaries are likely to match the users' needs well. In addition, MUSI accounts for user needs by offering a variety of meta-information, as will be discussed in more detail in Section 4.3.

To give a typical MUSI sample summary, let us assume that the user is interested in identifying papers on her preferred subject that she needs to access and read. She uses an IR system operating on a database of scientific papers in the medical domain.[4] The user retrieves documents by entering the query term "acetylcholine"[5] and opting for full meta-information. Each

[4] For MUSI the product LexiQuest Guide by LexiQuest is used for IR. The data were taken from the ESiA Online Journal of Anaesthesiology, mainly since it is available freely in both English and Italian on the Internet (`http://anestit.unipa.it/esiait/esiaing/esiaingfram.htm`).

[5] LexiQuest Guide allows for boolean queries.

```
PROP{ Value = P_ARG1_cause_ARG2;
      Time_Rep = [PRESENT, PRES_USUAL];
      Cat = V_SEN;
      Arg1 = PROP{ Value = P_antagonism_with_ARG1;
                   Cat = NP; Det = INDEF;
                   Arg1 = ITEM{ Value = C_acetylcholine;
                                Mod1 = [LOC, ITEM{
                                       Value = C_level;
                                       Det = DEF;
                                       Mod1 = [RESTR, ITEM{
                                               Value = C_sight;
                                               Number = PLUR; Det = DEF;
                                               Mod1 = [RESTR, C_muscarinic];
                                               Mod2 = [RESTR, ITEM{
                                                       Value = C_substance;
                                                       Number = PLUR;
                                                       Det = DEMONST1;}]; }]; }]; };
                   Mod1 = [RESTR, C_competitive]; };
      Arg2 = ITEM{ Value = C_effect;
                   Det = DEF; Number = PLUR; }; }
```

Figure 3: IRep4 Expression for “Die Wirkungen werden durch einen kompetitiven Antagonismus zu Acetylcholin auf dem Niveau der muskarinischen Bindungsstellen dieser Substanzen verursacht.” [The effects are caused by a competitive antagonism with acetylcholine on the level of the muscarinic sights of these substances.].

of the results has a “summarize” button associated with it. Pressing this button might yield, in a separate window, a text as shown in Figure 2.

3 Interlingua Semantic Representation

IRep4 forms the interlingua for cross-lingual processing in MUSI. Its core consists of a concept-based, hierarchical predicate-argument structure complemented by a rich variety of modifiers. Figure 3 shows a sample representation. In this section, the major features of IRep4 are described, and the relations between IRep4 elements and the German conceptual lexicon are explained.

The core interlingua elements are the lexical concepts, which are prefixed by P_ if they are predicative, i.e. they have arguments, and by C_ otherwise. Predicative concepts correspond to one or several lexemes in a language. Associated categorial information (Cat) suggests a word choice according to parts of speech: in the top of Figure 3, V_SEN indicates that the concept P_ARG1_cause_ARG2 has been realised as a transitive verb (“to cause”), whereas NP suggests the realisation of P_antagonism_with_ARG1 as a noun. Whenever nominal realisation is required, information about determination (Det) and number (Number) is mandatory.[6] Obviously, this information, too, is based on source-language analysis and need not transport to the target-language generation task in hand. But in many cases, it can guide generation safely. On the same basis, tense and aspect information (Time_Rep) is provided for sentence-valued elements.

Some linguistic phenomena and their representation in IRep4 are worth discussing in more detail. The sample representation corresponds to both the active and passive linguistic variant. A choice is made during generation based on the complexity of the arguments: if they differ largely, the shorter one is selected as subject; if it differs only marginally, active voice is preferred.[7]

[6] Default values apply for “no determiner” and “singular”, respectively.

[7] An empty ARG1, denoted by “ARG1 = 0;”, enforces a passive construction.

```
ITEM{ Value = C_Mandrake;
      Coref = I;
      Mod1 = [RESTR, PROP{
              Value = P_ARG1_known_for_ARG2;
              Time_Rep = [PRESENT, PERF_DURATION];
              Cat = V_SEN;
              Arg1 = ITEM{ Coref = I; };
              Arg2 = ITEM{ Value = C_property;
                           Number = PLUR;
                           Mod1 = [RESTR, C_mydriatic];
                           Mod2 = [LOC1, ITEM{
                                   Value = C_poss_pro; Coref = I; }]; };}]; }
```

Figure 4: An IRep4 Expression Showing Coreference: "Alraune, die für ihre pupillenerweiternden Wirkungen bekannt ist" [Mandrake, which is known for its mydriatic properties]

Prepositions are represented either as a part of subcategorised information in the lexicon entry for predicative concepts, such as `P_antagonism_with_ARG1`, or as a part of nominal lexicon entries relating to specific modifiers. For instance, the lexicon entry for `C_level`, corresponding to "Niveau" [level], associates the modifier `LOC` with the preposition "auf" [on]. If `C_level` appears as `Value` of a `LOC` modifier, the PP "auf dem Niveau" [on the level] can be generated.

The realisation of modifiers can depend on the part of speech of their head concept. Most prominently, the generic modifier `RESTR` may contain either a simple non-predicative concept, a structure `ITEM` containing a non-predicative concept, or a structure `PROP` containing a predicative concept (for the latter, see Figure 4).

In the first case, the concept corresponds lexically to an adjective, in which case a prenominal adjectival modifier will be generated. An `ITEM` structure containing a head concept that corresponds to a noun represents a generalised possessive, which normally corresponds to a postnominal genitive NP in German.[8] The third case subdivides again according to the lexeme chosen for the predicative concept. A verb induces a relative clause, whereas a noun again induces a generalised possessive. An adjective, in this case with filled arguments, either induces a relative clause with a copula, or a prenominal adjectival complex. Consider "der dem Arzt bekannte Fall" [the case known to the doctor – literally: "the to-the doctor known case"] and "der Fall, der dem Arzt bekannt ist" [the case that is known to the doctor]. Unlike the prenominal AP, the relative clause exhibits tense information, which is present tense by default. Usually, the decision whether to generate a relative clause or an AP depends on categorial information (`Cat = V_SEN` or `Cat = ADJP`).

Every realisation as a relative clause must exhibit a relation between a constituent of the relative clause and the noun the relative clause is attached to. The constituent in question is realised as the relative pronoun. It is identified by an `ITEM` structure containing only a feature called `Coref` that establishes a co-reference relation between different constituents of an IRep4 expression (cf. Figure 4). The co-reference relation is also used for possessive pronouns, as is also shown in Figure 4, and for personal pronouns. In all cases, appropriate features will be shared between the constituents generated from IRep4 elements with the same co-reference index in order to establish agreement in German.

IRep4 has been defined on the basis of a pre-existing suggestion by LexiQuest. Many additional features have been added [Chevreau *et al.*, 2000], and its semantics has been defined informally. As an interlingua, IRep4 is amenable to all analysis and generation components

[8] These require an article or an adjective. If neither is justifiable, a PP with the preposition "von" [of] is used instead. Compare "die Krankengeschichte des Mannes" [the case history of the man] with "Anzeichen von Vergiftung" [signs of toxicity].

involved. IRep4 is suitable for the representation of the semantics of very complex sentences, as they are often found in scientific documents. This does, however, not imply that the analysis components will always be able to come up with complete IRep4 representations. Resolution of proforms specifying `Coref`, identification of arguments and modifiers, and attachment ambiguities are the most prominent, and quite expected, sources of problems that may lead to fragmentary or incomplete output.

4 Generation Techniques Used in MUSI

We first introduce the system TG/2, which is re-used for the generation of German summaries, then present the generation from IRep4 expressions and finally turn to the generation of meta-statements.

4.1 An overview of TG/2

TG/2 [Busemann, 1996] is a shallow NLG system based on restricted production system techniques [Davis and King, 1977] that preserve the modularity of processing (the interpreter) and linguistic knowledge (the grammar), hence increasing transparency[9] and improving re-usability for NLG in various applications. In fact, TG/2 has been used before as a language generator

- in a multi-agent dialog system in the domain of appointment scheduling [Busemann *et al.*, 1994],
- for air quality report generation [Busemann and Horacek, 1998], and
- for situated agents suggesting activities in a conference scenario [Geldof, 2000].

These applications have in common the need for only a limited amount of linguistic coverage. They differ largely in the domain, the type of linguistic constructions needed, and the granularity required for language modelling. For instance, the appointment scheduling domain required a fine-grained account for temporal expressions, whereas the air quality report system was very coarse-grained, but covered a much more diverse spectrum of linguistic constructions.

A different language representing the input had to be designed for each of these applications. To encode input language expressions, a generic feature structure formalism is used that is interpreted by various constraint processing systems. With help of a grammar, TG/2 relates pieces of input structures to surface strings. If the input languages contain task and domain concepts rather than linguistic semantic expressions, TG/2 does more than traditional realisation components: it covers all language NL-relevant aspects of the NLG process. In a production system, deliberate planning and complex interdependencies cannot be encoded. The applicability conditions for production rules would become overly complex and opaque. As a consequence, TG/2 shortcuts sentence planning tasks such as aggregation, lexical choice and the generation of referring expressions considerably. This is achieved by using templates and canned text whenever this is justified by the task and the domain in hand. Shallow generation *extends* the information contained in the input (by virtue of canned text), whereas shallow analysis *ignores* information contained in the input text. Applications relying on deliberate sentence planning can probably not be implemented within TG/2, but require an extra module that feeds TG/2.

Due to the diversity of input languages and the domain-specific, though modest, linguistic coverage required, most of the linguistic knowledge was built from scratch each time. This required less effort than adapting existing broad-coverage resources such as KPML [Bateman, 1997] or SURGE [Elhadad and Robin, 1996] would have (cf. [Busemann and Horacek, 1998]). The flexibility to adapt grammars to different levels of granularity turned out to be crucial for

[9] Most importantly, a production is not allowed to modify the conditions deciding about the applicability of other productions as a side effect.

```
(defproduction sentence "simplesample"
  (:PRECOND (:CAT S
             :TEST ((subject-top-p)))
   :ACTIONS (:TEMPLATE  "Welcome to the SimpleSample test system: &newline;"
                        (X1 :RULE NP (get-param ’subject))
                        (X2 :RULE VP (get-param ’predicate))
                        "."
             :CONSTRAINTS (X1.NUMBER = X2.NUMBER))))
```

Figure 5: A TG/2 Rule Defining a Sentence Template.

successfully reusing TG/2. It is achieved by integrating canned text, templates, and context-free rules into a single formalism. A coarse-grained model relies on many canned text parts; a fine-grained model uses many rules to model a particular phenomenon. The generation process uses the rules to create a derivation tree, the leaves of which are used to create the system output. Every rule has a context-free backbone (cf. S → NP VP in Figure 5). A rule is always applied in a context of a current category and a piece of input structure.[10] Productions are applied through the well-known three-step processing cycle:

1. identify the applicable rules,
2. select an applicable rule, and
3. apply that rule.

A rule is applicable if its preconditions are met. These involve matching the rule's category with the current category and a set of programmable test predicates on the piece of input structure. For instance, in Figure 5, `subject-top-p` checks for information indicating that the sentence should start with the subject. An applicable rule is selected on the basis of some freely programmable conflict resolution mechanism. A rule is applied by carrying out its actions in a top-down, depth-first and left-to-right manner. A TG/2 rule is successfully applied if all actions are carried out. The rule returns the concatenation of the substrings produced by the "template" actions. If an action fails, backtracking can be invoked flexibly and efficiently using memoisation techniques (for details see [Busemann, 1996]).

The actions of a rule include the activation of other rules (`:RULE`) or the return of an ASCII string as a (partial) result. When selecting other rules by virtue of a category, the relevant portion of the input structure for which a candidate rule must pass its associated tests must be identified. The access function `get-param` in Figure 5 yields the substructure of the current input depicted by the argument. The first action causes `NP` to be the new current category; the relevant substructure is the information encoded under `subject` in the input. The cycle is repeated. Processing terminates when all actions of the first rule have produced a terminal yield. Note that this implies that grammars may have recursive rules, as termination is eventually guaranteed by the finiteness of the input.

Upon termination, the terminal yield of the derivation tree is read off as a sequence of strings or morphologically annotated word stems. The latter uniquely correspond to fully inflected word forms, which can be produced either from a full-form lexicon or with the help of a morphological inflection component such as Morphix-3 with its large and extensible stem lexicon for German (over 100.000 entries) [Finkler and Neumann, 1988]. This yields a sequence of strings the concatenation of which forms the final output of TG/2.[11]

[10] At the beginning a dedicated start category is used.

[11] Actually, TG/2 offers different output formats if a set of specific markup signs is used in the grammar. Parsing the result string and replacing the marked parts contained in it according to some output parameter produces HTML, plain ASCII or LaTeX format.

Agreement relations are encoded into TGL by virtue of a PATR style feature percolation mechanism [Shieber *et al.*, 1983]. The rules can be annotated by equations that either assert equality of a feature's value at two or more constituents, or introduce a feature value at a constituent. The constraint in Figure 5 requires the categories NP and VP to agree in number. This general mechanism provides a considerable amount of flexibility and goes beyond simple template filling techniques.

4.2 Generation of summary content

4.2.1 The protogrammar

In TG/2, the relations between the input and the grammar are tight. It seems that for any new input language, a new grammar must be designed. On closer inspection, however, the parts of a grammar rule depending on input elements can be isolated and treated as an interface between the grammar and any input language. If an input language changes, the test predicates and the access functions need to be recoded. This interface allows us to develop generic grammar knowledge that abstracts from specific semantics of test predicates and access functions. We call such a generic grammar a *protogrammar*, as it is supposed to form the reusable basis for different instances geared towards different applications. Technically, a protogrammar can be instantiated by defining the test predicates and access functions needed for the input language in question.

The MUSI project with its comparatively fine-grained IRep4 expressions calls for the modelling of a generic reusable resource that abstracts from IRep4-specific interfaces. In practical project work, NLG grammar development cannot be postponed until the results of analysis become available. With the above consideration in mind, it was possible to start even before a precise definition of IRep4 was available.

The protogrammar covers the main types of sentential structures, as specified by the Duden grammar [Dudenredaktion, 1998]. The NP syntax comprises prenominal APs (on the basis of adjective subcategorisation frames), generic possessive constructions, a temporal, a locative and an adverbial modifier and a relative clause. In addition, nouns and adjectives can subcategorise for specific arguments.

How could a protogrammar be developed independently of a particular input language like IRep4, as it needs testing? One idea consists in having an empty input, no tests on it, and no access functions. With terminal elements in the grammar representing the parts of speech instead of the words, TG/2 would then enumerate the admissible syntactic structures. Unfortunately, it would not terminate on recursive rules. Hence it is mandatory to have a simple, intuitive input representation that could either be easily adapted to IRep4 later, or onto which IRep4 expressions could be mapped. Basically, this representation would determine the depth of the nesting of constituents, specify morpho-syntactic features such as case, number, tense etc., indicate the prepositions and distinguish the syntactic adjuncts at the sentence and NP level. This new, intermediate representation is called GIL (for generator-internal language).

4.2.2 Architectural considerations

The GIL layer is independently justified by the observation that feeding IRep4 expressions into TG/2 would be impossible for several reasons. First, the formalisms of IRep4 and TG/2 are incompatible, and second, such a mapping would involve very complicated test predicates and access functions. A closer look at the MUSI application made it clear that the amount of sentence planning needed could not very easily be mastered within the paradigm of production rules, i.e. within TG/2. Due to the source language being different from German, IRep4 expressions may reflect properties that cannot be realised straightforwardly in German. For instance, Italian allows for several heavy-weight post-nominal APs or gerunds, which in German would

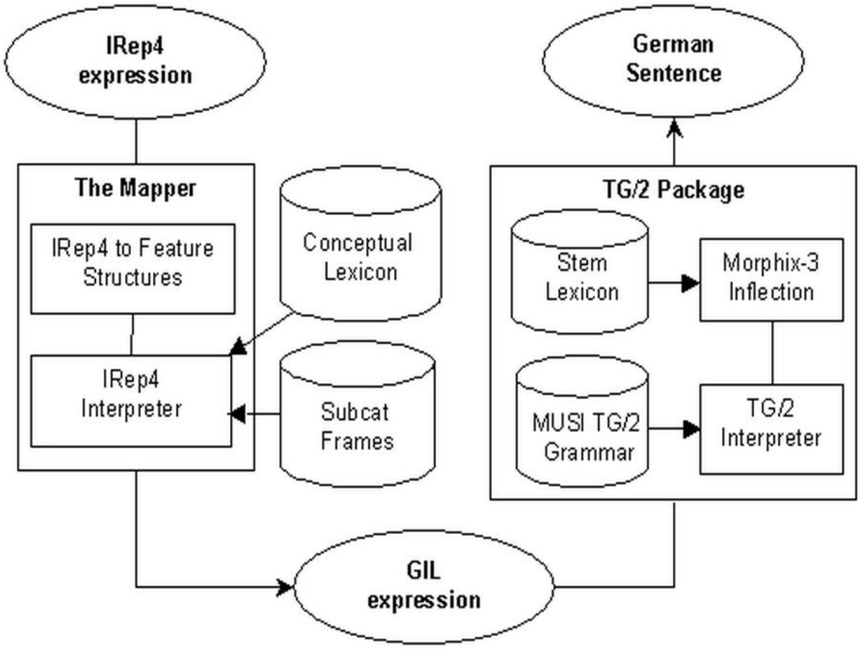


Figure 6: The Architecture of the MUSI Generator for German.

be realised as prenominal APs or a relative clause. Another problem is the generation of articles from English-based IRep4 expressions: often no article is available where an indefinite article is mandatory in German. A fourth difficulty arose from some information not being represented at all in IRep4, such as constituent order. We use heuristic information to decide between e.g. active and passive voice.

To deal with these problems, a new component was built, which was called quite vaguely the *Mapper*. In its first stage it translates IRep4 objects into TG/2 feature structures. This is always feasible, as IRep4 is a tree language. For Coref, the lexical realisation of the referent concept is looked up in order to specify stem, number, gender and person information at a unique location in the feature structure. This information is used later for the generation of pronouns.

The second stage of the Mapper is dedicated to sentence planning tasks:

Lexical choice: Concepts must be related to target-language words. While this could, in principle, be achieved by TG/2 grammar rules of the kind "if concept is X, then word is Y", the conflict set of the respective word classes could become extremely large. In other words, every time, TG/2 generated a noun, it would have to execute the test predicates of all noun rules. By assigning lexical choice to the Mapper, this effort can be saved. As a consequence, GIL contains German words and no concepts. Moreover, the subcategorisation information and the lexical relations between types of modifiers and prepositions outlined in Section 3 could not be represented at the lexical level in TG/2, as TG/2 processes top-down and would fail to access that information when it is needed.

Syntactic choice: The IRep4 objects must be mapped onto sentence-syntactic structures. TG/2 encodes many different sentence structures, each in different word orders and as main or subclauses. While each rule could have its applicability tests executed on the IRep4 input, this again amounts to a huge overhead that can be avoided by reflecting subcategorisation information in GIL through the name of a suitable sentence plan that must fit the tests associated with the sentence rules of the grammar. The prepositions and the syntactic case they inflict on the governed NP are also encoded in GIL (cf. Figure 7).

```
[(SENTENCE DECL)
 (VC [(SBP S2)         ;;name of sentence plan
      (G AKTIV)         ;;active voice
      (STEM "verursach")])
 (DEEP-SUBJ [(TOP Y)         ;;this constituent to the fore-field
             (DET INDEF)         ;;indefinite article
             (NR V2)         ;;name of nominal plan
             (STEM "antagonismus")
             (PP-ATR [(MODALITY
                        [(PP-OBJ
                           [(TERM [(DET DEF)         ;;definite article
                                   (STEM "bindungsstelle")
                                   (ADJ [(STEM "muskarinisch")
                                         (DEG POS)])
                                   (TERM [(DET DEMONST1)         ;;demonstrative
                                          (STEM "substanz")])])])
                            (STEM "Niveau")
                            (DET DEF)
                            (PREP AUF-DAT)])])         ;;P governs dative NP here
                      (STEM "acetylcholin")
                      (DET WITHOUT)         ;;no article
                      (PREP ZU)])         ;;this P always governs dative NP
             (ADJ [(STEM "kompetitiv")
                   (DEG POS)])])
 (DEEP-AKK-OBJ [(DET DEF)
                (STEM "wirkung")])]
```

Figure 7: A GIL Feature Structure for "Ein kompetitiver Antagonismus zu Acetylcholin auf dem Niveau der muskarinischen Bindungsstellen dieser Substanzen verursacht die Wirkungen." [The effects are caused by a competitive antagonism with acetylcholine on the level of the muscarinic sights of these substances.]. Comments are separated by semicolons. The structure corresponds to that of Figure 3. It is simplified by omitting gender, number, mood and phrase type information.

Given these tasks, the output of the Mapper, GIL, has grown into much more than the minimal input structure that was required for protogrammar testing. It reflects the decision to keep both lexical and syntactic choice outside TG/2.

The Mapper methods taken together keep the conflict sets in TG/2's derivation process reasonably small. For a sentence plan, some eight to ten rules need to be considered. Usually, these describe word order variations: in German declarative clauses, any phrasal constituent can occupy the first position, and yes-no questions and subclauses require a different position of the finite verb (first and last position, respectively).

The overall generation architecture is reflected in Figure 6. It consists of two components, the Mapper and the TG/2 package. The Mapper first creates feature structures that are then interpreted using the conceptual lexicon and the information about subcategorisation to yield GIL representations. The TG/2 package then realises GIL expressions with help of the MUSI grammar, an instance of, and extension to, the protogrammar. Words are inflected using the Morphix-3 component with its stem form lexicon of German.

4.2.3 Lexical and syntactic choice

This section describes the major aspects of lexical and syntactic choice (more details can be found in [Busemann *et al.*, 2001]).

Non-predicative concepts correspond to nouns, adjectives or adverbs. For nouns, the conceptual lexicon represents information about stem and gender. In addition, nominal entries contain the preposition that the noun induces if it is the head of a particular modifier (cf. Section 3 and below). The concepts are chosen in such a way that German nominal compounds are always based on single concepts, thus avoiding the necessity of a compound generator to produce, for instance, A-N compounds such as "Ernährungsungleichgewicht" [nutritional imbalance] or N-N compounds such as "Muskelfaserzelle" [muscle fiber cell].[12]

For adjectives, the conceptual lexicon encodes the stem and a marker on whether the word is used as a past participle or an adjective proper. This information is merged together with determiner, quantifier and number information into a GIL term (see Figure 7).

Possessive and personal pronouns are signalled by concepts, e.g. C_poss_pro. The Mapper generates a GIL representation of the type of pronoun and associates with it the morpho-syntactic information for gender, number and person, which is either specified explicitly in IRep4 or derived from the Coref feature associated with the proform.

Predicative concepts can be realised as verbs, nouns or adjectives. The conceptual lexicon may have multiple entries for these concepts, each of which points to a specific subcategorisation frame that is used to generate the GIL expression. Gender information is added for nominal entries, and verbal as well as adjectival entries are associated by a preposition in the case one of their arguments is realised as a PP. If the preposition may induce different surface cases on its NP, the correct case is encoded as well (cf. Figure 7).

For instance, the lexicon entry for the concept P_Arg1_treat_Arg2_with_Arg3 below has a verbal and a nominal realisation in German. The verbal reading is attached with the subcategorisation frame vr13, and the nominal one with nr3. These frames account for different combinations of empty or filled arguments. If e.g. Arg1 in the verbal frame is empty, a passive construction is initiated here.

```
(add-P2lex 'P_Arg1_treat_Arg2_with_Arg3
           '(vr13 nr3)
           '(($stem-v . "behandel")
             ($prep-pp . mit)
             ($stem-n . "Behandlung")
             ($gender . fem)))
```

The following feature structure reflects the main properties of vr13. The material under out is used in GIL. The feature trans specifies input and output of the mapping process for the arguments under in.[13] It uses the material defined under out by virtue of feature co-reference[14]. According to trans, the value of arg1 is realised in GIL as an NP under deep-subj, the value of arg2 as an NP under deep-akk-obj and the value of arg3 as a PP under pp-obj. The values prefixed by $ correspond to the respective lexical information, as explained above. Here, we get "behandel" as the stem and "mit" as the preposition.

```
[(in [(arg1 %1)
      (arg2 %2)
      (arg3 %3)
      (cat {v_sen, adjp})])
 (out [(sentence decl)
       (vc [(stem $stem-v)
            (sbp s13)])
       (deep-subj %4 = [(cat np)])
       (deep-akk-obj %5 = [(cat np)])
```

[12] Obviously a compound generation component would have added much flexibility to concept-based text generation of German.

[13] < and > denote list delimiters.

[14] Co-reference is encoded by designators starting with %. Identical designators denote identical feature structure objects.

```
      (pp-obj %6 = [(cat pp) (prep $prep-pp)])])
 (trans < [(in %1) (out %4)], [(in %2) (out %5)], [(in %3) (out %6)] >)]
```

The subcategorisation frame also restricts the set of applicable grammar rules by specifying the feature `sbp`, which is checked by TG/2 test predicates. In the example, only TG/2 grammar rules passing the test on `s13` will be applicable.

If several lexical realisations are possible, a deliberate choice must be made. The choice may be restricted by constraints imposed by a governing predicative concept, forcing arguments to be encoded as e.g. an NP, a PP or a sentence (cf. the example above). This specification overrides source-language-based IRep4 specifications for `Cat` that are otherwise used as a guide.

The filled frame must be compatible with the complete IRep4 input. Otherwise, an alternative frame is selected according to the lexicon entry. If the frame fits, the mapping is applied on each of the arguments recursively. Eventually the modifiers are mapped recursively as well.

A modifier is a pair consisting of the modifier type and an atomic or complex structure representing the modifier's content. The word selected for the head concept determines whether the modifier will be realised as a word (e.g. an adjective), a phrase (e.g. an NP) or a sentence. Modifiers may also be realised as PPs in which case the preposition is derived from the lexicon entry of the head, or - in case this fails - by a default preposition associated with the type of the modifier.

4.2.4 The MUSI grammar

In view of the documents under investigation, the protogrammar was instantiated and extended as a MUSI-specific grammar comprising all possible constructions occurring in the sentences that could ever be selected for deep analysis. The resulting MUSI grammar comprises about 950 rules with 135 categories, 134 test predicates[15], 8 access functions, 9 string-valued functions and 14 features for constraints.

Grammar development took place with the special-purpose editor eGram that enforces a consistent way of defining grammar objects by allowing the definition of complex objects only if all elements are defined. For instance, for the definition of a rule, the categories, test predicates, access functions and constraints used must be defined first. eGram provides the grammar writer with different views of the grammar and its components, thus keeping even a large set of rules manageable. At the same time, it abstracts from implementation-specific formats used for TG/2 grammars. The results of editing are transformed into the format processed by TG/2 through an XML interface.

Figure 8 provides a screenshot of eGram, showing a TG/2 rule applicable to the GIL expression in Figure 7. Note that flexibility is gained by defining optional elements on the right-hand side of the rule (`:OPTRULE`) that are ignored if no input for them is retrieved by the respective access function.

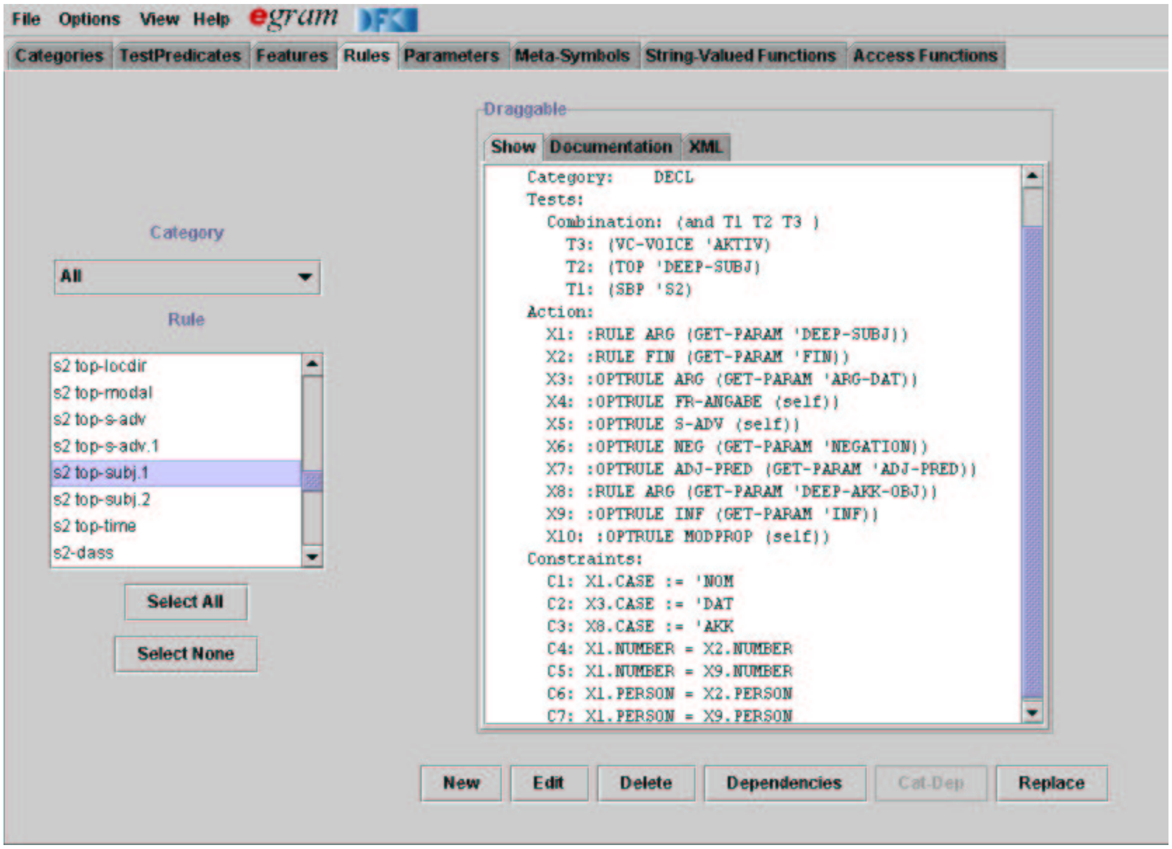
4.3 Generation of meta statements

Due to the need of deep linguistic processing for cross-lingual summarisation, MUSI summaries may be complemented by additional statements about the text not contained in the original. This option is not available in its full flexibility with summarisation techniques that are based on sentence selection only.

We call statements *about* the document *meta statements*. Meta statements are categorized according to types of information, enabling the user to select or de-select them individually.

Another important option offered by the types consists in computing their order within the document structure. Meta statements may come at the beginning or at the end of the summary,

[15] Note that the number of test predicates could be reduced drastically by adding IRep4-specific knowledge about feature paths.



```
Category:    DECL
Tests:
  Combination: (and T1 T2 T3 )
    T3: (VC-VOICE 'AKTIV)
    T2: (TOP 'DEEP-SUBJ)
    T1: (SBP 'S2)
Action:
  X1: :RULE ARG (GET-PARAM 'DEEP-SUBJ))
  X2: :RULE FIN (GET-PARAM 'FIN))
  X3: :OPTRULE ARG (GET-PARAM 'ARG-DAT))
  X4: :OPTRULE FR-ANGABE (self))
  X5: :OPTRULE S-ADV (self))
  X6: :OPTRULE NEG (GET-PARAM 'NEGATION))
  X7: :OPTRULE ADJ-PRED (GET-PARAM 'ADJ-PRED))
  X8: :RULE ARG (GET-PARAM 'DEEP-AKK-OBJ))
  X9: :OPTRULE INF (GET-PARAM 'INF))
  X10: :OPTRULE MODPROP (self))
Constraints:
  C1: X1.CASE := 'NOM
  C2: X3.CASE := 'DAT
  C3: X8.CASE := 'AKK
  C4: X1.NUMBER = X2.NUMBER
  C5: X1.NUMBER = X9.NUMBER
  C6: X1.PERSON = X2.PERSON
  C7: X1.PERSON = X9.PERSON
```

Figure 8: A Screenshot of eGram, a Special-Purpose Grammar Editor for TG/2.

and some types can even be interspersed with the content summary (e.g. information about the source of content information). The user can also select, e.g. from a menu, which meta information should come at the outset, within, or at the end of a summary. Unfortunately, this option is not fully operational within the MUSI system. Figure 2 has all meta statements in front of the summary content.

Meta statements are generated in a straight-forward manner using templates, as the wording can be designed a priori. As a consequence, the choice of stylistic variations for them resides with the designer of the summary. This is different for the document content, where the results of analysis and summarisation largely restrict the freedom of choice.

Type1: Information about source language.

- " The source document is written in English."
- Information required: source language.
- Available from analysis.

Type 2: Information about query relevance.

- "It is relatively relevant to the query."
- Information required: query relevance.
- Available from query expansion and analysis.

Type 3: Information about the document structure, size and length. We may specify whether there was an abstract, whether there is any section structure not reflected in the summary structure, and the number of tables and figures.

```
[ (LANGUAGE english)
  (Q-REL quite-rel)
  (DOC-INFO [ (GENERAL [ (TABLES 1) (FIGURES 3) (BIBL T)
                         (APPENDICES 0) (WORDS 3454)])
              (SUMMARY [ (CONCENTRATE < [ (SECTION-ID discussion)
                                          (SENTENCES all) ] >)
                         (MULTIPLES < > ) ]) ]) ]
```

Figure 9: The Input for the Generation of the Meta-Statements in Figure 2. LANGUAGE corresponds to Type 1, Q-REL to Type 2, DOC-INFO.GENERAL to Type 3, and DOC-INFO.SUMMARY to Type 4 meta statements.

- "The article is 3454 words long, it has one table, three figures, and a bibliography."
- Information required: top-level document structure (abstract, number of sections, conclusion, appendices), number of words, tables, figures, etc.
- Available from XML tags.

Type 4: Source of the content information within document. For each sentence or section in the summary, the part of the original document it is summarising is described. A single statement is used when it is simple and concise. A more generic statement is used if details are complicated (e.g. if every structure relates to a different element of the document structure).

- "The first three sections of the summary are based on the abstract. The section on Results is entirely based on the first paragraph of Section 2." — "The summary comes from various parts of the document. The section on Results is based on the Abstract and the Conclusion."
- Information required: Association between IRep4 objects and elements of the document structure (any level; "first paragraph of Section 2" is much more specific than "Conclusion").
- Available from XML tags.

It would be interesting to identify repetitive statements in the document within the extraction module, as these may be particularly relevant (cf. path DOC-INFO.GENERAL.MULTIPLES in Figure 9.

Figure 9 shows an input representation underlying the sample summary in Figure 2. It is written in the same feature structure formalism as GIL, though entirely different in content. Rather than specifying linguistic structures, it defines semantic content in a non-linguistic way. The information underlying it is compiled during the summarisation process. This representation style very closely resembles the input representations for shallow TG/2 applications, e.g. the air quality report system [Busemann and Horacek, 1998].

The template generation rules used in TG/2 for meta statements are much more coarse-grained than those dedicated to GIL. They use portions of canned text, complemented by variables corresponding to the specifications in the input.

Both kinds of rules are written in the same formalism and are part of the same grammar.

5 Conclusions and Future Directions

This paper described NL generation techniques used for cross-lingual summarisation of scientific medical texts. The approach chosen is based on source-language sentence extraction, creation of an interlingua representation of the extracted sentence, and target language generation from the interlingua. The MUSI system is fully implemented and serves as a demonstrator of the feasibility of the approach.

A distinguishing feature of MUSI is the usage of meta-information to tailor the summaries towards the user's needs. Crossing the language barrier opens up the possibility of including text that was not part of the original document. This opportunity was only exploited to a small degree in the work reported here. Including knowledge about the user's interest regarding content could cause all components of the system to extend the general relevancy criteria used so far. Uninteresting material could be ignored, and interesting parts highlighted in the summary.

The linguistic resources used had to be adapted to the domain. To keep this effort manageable, only a small subset of documents has been used. Scaling the resources up to a full-fledged application with a representative coverage of the domain in hand would require considerable support regarding the creation of lexical resources. The use of domain thesauri is certainly one of several possibilities, depending on their availability in the requested languages.

The degree of grammatical coverage required is very ambitious given the elaborate style of medical text. Automatically distinguishing relevant from less relevant information would reduce the grammatical complexity considerably. Less relevant information is often encoded in parenthetical expressions and restatement. The concepts used do not belong to medical terminology (consider e.g. "in the clinical case described below"). Such information can be detected by the lexicon access component and by the syntactic parser.

The generation grammars are likely to become stable fairly quickly during a scaling-up process. However, several sentence planning problems due to language-specific representations at the interlingua level could not be solved satisfactorily. They call for extensions to the system, and new problems at that level are likely to occur if the system is scaled up. Future work on generation should thus concentrate on designing some kind of "transfer knowledge" that explicitly deals with the language-specific reminders in the interlingua. For each language pair, a relatively small set of rules depending on the source and target languages could guide the sentence planner in creating stylistically better GIL representations.

In conclusion, the results reported here represent a first step towards cross-lingual text summarisation. It is still a long way to reach a viable application system.

Acknowledgments

This work has been partially supported by the European Commission under contract MLIS-5015. Besides DFKI, the MUSI Consortium includes LexiQuest (LQ, Paris, France) and the Istituto Linguistica Computazionale (ILC, Pisa, Italy). Part of the R&D work described in Sections 2 and 3 has been carried out by researchers at LQ and ILC, most notably Roberto Bartolini (ILC), Nicoletta Calzolari (ILC), Emmanuel Cartier (LQ), Karine Chevreau (LQ), José Coch (LQ), and Alessandro Lenci (ILC). I am grateful to Ana Água, Christian Au, Matthias Großkloß, Matthias Rinck and Tim von der Brück (all DFKI) for their implementation work in the context of the MUSI generation system for German. I remain responsible for any misconceptions and mistakes that may be ontained in this contribution.

References

[Bateman, 1997] John Bateman. KPML delvelopment environment: multilingual linguistic resource development and sentence generation. Report, German National Center for Information Technology (GMD), Institute for integrated publication and information systems (IPSI), Darmstadt, Germany, January 1997. Release 1.1.

[Busemann and Horacek, 1998] Stephan Busemann and Helmut Horacek. A flexible shallow approach to text generation. In Eduard Hovy, editor, *Nineth International Natural Language Generation*

Workshop. Proceedings, pages 238–247, Niagara-on-the-Lake, Canada, 1998. Also available at `http://xxx.lanl.gov/abs/cs.CL/9812018`.

[Busemann *et al.*, 1994] Stephan Busemann, Stephan Oepen, Elizabeth Hinkelman, Günter Neumann, and Hans Uszkoreit. COSMA–multi-participant NL interaction for appointment scheduling. Technical Report RR-94-34, DFKI, Saarbrücken, 1994.

[Busemann *et al.*, 2001] Stephan Busemann, Ana Água, Christian Au, Matthias Grosskloss, and Matthias Rinck. MUSI text generation. report on functionality and performance (Part I: German generation). Technical report, MUSI project deliverable D5.1, 2001.

[Busemann, 1996] Stephan Busemann. Best-first surface realization. In Donia Scott, editor, *Eighth International Natural Language Generation Workshop. Proceedings*, pages 101–110, Herstmonceux, Univ. of Brighton, England, 1996. Also available at the Computation and Language Archive at `http://xxx.lanl.gov/abs/cmp-lg/9605010`.

[Chevreau *et al.*, 2000] Karine Chevreau, José Coch, and Emmanuel Cartier. MUSI internal text representation specification. Technical report, MUSI project deliverable D2.1, 2000.

[Davis and King, 1977] Randall Davis and Jonathan King. An overview of production systems. In E. W. Elcock and D. Michie, editors, *Machine Intelligence 8*, pages 300–332. Ellis Horwood, Chichester, 1977.

[Dudenredaktion, 1998] Die Dudenredaktion. *Duden. Die Grammatik. Grammatik der deutschen Gegenwartssprache*, volume 4 of *Duden - Das Standardwerk zur deutschen Sprache*. Dudenverlag, Mannheim - Wien - Zuerich, 6. edition, 1998.

[Elhadad and Robin, 1996] Michael Elhadad and Jacques Robin. An overview of SURGE: a reusable comprehensive syntactic realization component. In Donia Scott, editor, *Eighth International Natural Language Generation Workshop. Demonstrations and Posters*, pages 1–4, Herstmonceux, Univ. of Brighton, England, 1996.

[Finkler and Neumann, 1988] Wolfgang Finkler and Günter Neumann. Morphix: A fast realization of a classification–based approach to morphology. In H. Trost, editor, *Proceedings der 4. Österreichischen Artificial–Intelligence Tagung, Wiener Workshop Wissensbasierte Sprachverarbeitung*, pages 11–19, Berlin, August 1988. Springer.

[Geldof, 2000] Sabine Geldof. From context to sentence form. In *Proc. 1st International Language Generation Conference*, 2000.

[Mani and Maybury, 1999] Iderjeet Mani and Mark T. Maybury, editors. *Advances in Automatic Text Summarization*. MIT Press, 1999.

[Radev and McKeown, 1998] Dragomir Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, September 1998.

[Shieber *et al.*, 1983] Stuart Shieber, Hans Uszkoreit, Fernando Pereira, Jane Robinson, and Mabry Tyson. The formalism and implementation of PATR-II. In Barbara J. Grosz and Mark E. Stickel, editors, *Research on Interactive Acquisition and Use of Knowledge*, pages 39–79. AI Center, SRI International, Menlo Park, CA., 1983.